UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

APPLICATIONS OF THE GB2 FAMILY OF PROBABILITY DISTRIBUTIONS IN COLLECTIVE RISK THEORY*

by

J. David Cummins[1]
Georges Dionne[2]
and
Lionel Maistre[3]

[1] University of Pennsylvania.
[2] Département de science économique et C.R.T., Université de Montréal.
[3] Université de Montréal.

* This article is preliminary and may not be reproduced or quoted without the permission of the authors. The Huebner Foundation and C.R.S.H. (Canada) provided financial support to Georges Dionne. The research assistance of Marie-Josée Fortin is greatly appreciated.

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Université de Montréal
Centre de recherche sur les transports - Publication #532
Département de Science économique - Cahier #8719
May 1987

Dépôt légal - 3e trimestre 1987

ABSTRACT

This article proposes the family of probability distributions known as the generalized beta of the second kind (GB2) as a principal set of distributions for modeling insurance loss processes. The GB2 family encompasses many commonly used distributions such as the log-normal, gamma, and Weibull. It also includes distributions such as the Burr and generalized gamma which have significant potential for improving the distributional fit in many applications involving heavy-tailed distributions. Most members of the GB2 family can be generated as mixtures of well-known distributions, thus facilitating theoretical modeling of claims from heterogeneous populations. An example is presented which involves fitting the log-gamma and log-Burr distributions to a sample of fire claims. The results suggest that seemingly slight differences in modelling the tails of severity distributions can lead to substantial differences in reinsurance premiums and quantiles of simulated total claims distributions.

Keywords: Generalized beta of the second kind, Insurance losses, Gamma, Burr, Fire claims.

RESUME

Cet article suggère la famille de distributions connue sous le nom de beta généralisée du deuxième type (GB2) comme ensemble de distributions pouvant modéliser les pertes assurées. La famille de distributions proposée a comme cas particuliers la lognormale, la gamma, et la Weibull. Elle comprend également comme cas particuliers la Burr et la gamma généralisée connues comme étant des distributions avec queues épaisses. La plupart des membres de la famille de distributions proposée peuvent être obtenues par des combinaisons de distributions bien connues, facilitant ainsi la modélisation théorique des réclamations d'assurance de populations hétérogènes. Nous présentons un exemple où nous ajustons les distributions log-gamma et log-Burr à un échantillon de réclamations d'incendie. Les résultats suggèrent que de petits écarts dans la modélisation des queues des distributions de pertes peuvent générer des différences importantes dans les primes de réassurance et dans les quantiles des distributions de réclamations simulées.

Mots-clés: Beta généralisée du deuxième type, pertes assurées, Gamma, Burr, réclamations d'incendie.

APPLICATIONS OF THE GB2 FAMILY OF PROBABILITY DISTRIBUTIONS IN COLLECTIVE RISK THEORY

The choice of probability distributions to model loss processes in risk theory often has been dictated by considerations of mathematical tractability. Two dimensions of tractability are important -- parameter estimation and tabulation of the distribution function. The lognormal distribution, for example, has been used frequently at least in part because maximum likelihood estimators (MLEs) of the parameters are easy to compute and because tables of the normal distribution are readily available. The Pareto has easily estimable MLEs and a closed-form distribution function. For the gamma (Pearson Type III), method of moments parameter estimates are easy to obtain; and tables of the distribution function are available, if somewhat less ubiquitous than tables of the normal distribution. Other distributions such as the log-t have been utilized for similar reasons (see Cummins and Freifelder [1978]).

While the traditional distributions have often provided an adequate description of insurance data, the limitations imposed by tractability on the feasible set of distributions implies that the "best" distribution will not necessarily be selected in any given application. Fortunately, recent advances in computing have greatly increased the range of distributions accessible to the risk theorist. For example, Paulson has made extensive use of the stable family of distributions, which includes as special cases the one-tail Pareto, the normal, and the Cauchy, among others (see, for example, Paulson and Faris [1985]). Aiuppa [1986] utilizes a computer program that can estimate parameters and compute percentile points of distribution functions for any member of the Pearson family. This includes the Pearson Type IV, which statisticians traditionally have considered intractable (see, for example, Johnson and Kotz [1970, p. 12]). Finally, McDonald [1984] and McDonald and Butler [1985]

describe economic applications of the family of distributions known as the *generalized beta of the second kind* or *GB2 distributions*.

The purpose of this article is to propose the GB2 family as a principal family for use in modeling insurance loss distributions and to illustrate its use in describing a set of fire loss data. The paper goes beyond McDonald and McDonald and Butler by using individual claim data. It represents an application of the mean residual life (MRL) methodology suggested by Hall and Wellner [1981] and Hogg and Klugman [1984]. An important finding is that commonly-used distributions such as the log-gamma and log-normal may significantly underestimate the tails of insurance claim data.

The GB2 Family of Distributions

Unlike the generalized beta of the first kind (GB1), which is mentioned in risk theory texts (see, for example, Buhlmann [1970]), the GB2 has no upper limit and hence is likely to be applicable for severity distributions and other risk theory applications where the upper tail has no theoretical boundary.[1] The density function of the GB2 distribution is given in equation (1):

$$(1) \qquad f(x;a,b,p,q) = \frac{|a|x^{ap-1}}{b^{ap}B(p,q)[1+(x/b)^a]^{p+q}} \quad , \; x > 0$$

where B(p,q) = the beta function. Moments of the distribution (about the origin) are given by:

$$(2) \qquad E(x^h) = \frac{b^h B(p+h/a, q-h/a)}{B(p,q)}$$

The distribution function is:

$$(3) \qquad F(x) = \frac{\left(\frac{(x/b)^a}{1+(x/b)^a}\right)^p}{pB(p,q)} \; {}_2F_1\left[p, 1-q, \frac{(x/b)^a}{1+(x/b)^a}, p+1\right]$$

where ${}_2F_1[c,d,z,e]$ = the generalized hypergeometric series (see Abramowitz and Stegun [1965]).

Although it is possible to tabulate the distribution function using equation (3), it is often more convenient to use numerical integration. Either approach can be

used on a personal computer, given the appropriate software. If either a computer program or a table is available for the F distribution, percentile points of the GB2 can be obtained from the F by utilizing the transformation: $y = (q/p)(x/b)^a$. If x is GB2, y will be F with degrees of freedom $\nu_1 = 2p$ and $\nu_2 = 2q$.

The importance of the GB2 distribution for risk theory is that it has great flexibility due to the availability of four parameters. In addition, it encompasses many of the traditional distributions as special cases. Among these are the gamma, generalized gamma, lognormal, Burr, exponential, and Weibull. A partial distribution tree is presented in Figure 1.[2] For even more general distribution trees, see McDonald [1984] and McDonald and Butler [1985].

Members of the GB2 family arise as mixtures of a wide range of underlying distributions. For example, the log-t has been shown to be a lognormal mixed with an inverse gamma (see Cummins and Freifelder [1978] and Hogg and Klugman [1984]). The Burr arises as a Weibull distribution mixed with a gamma. Other mixture distributions leading to members of the GB2 family are presented in McDonald and Butler [1985]. Thus, the GB2 family has a theoretical justification as a representation of claims arising from a heterogeneous population of exposures.

Applying the GB2 Distribution

In applying the GB2, it is generally advisable to start with the simplest members of the family and to move up the distribution tree until no further gains in the fit of the distribution to the empirical data are obtained. Estimation can be carried out using either the method of moments or maximum likelihood estimation. Even the former technique is likely to require a computer for distributions lying on higher levels of the tree. The most general member of the family requires the solution of four nonlinear simultaneous equations to obtain the moments.

Fortunately, a number of programs are available to perform these tasks. Among the available programs are: (1) The International Mathematics and Statistics Library (IMSL [1985]), a set of Fortran subroutines that includes a nonlinear simultaneous equation solver; (2) Statistical Analysis System (SAS

Institute [1985]), a package of statistical programs that is widely used in economics; and (3) Gauss (see Edlefson and Jones [1986]), a math/stat program for IBM compatible personal computers. Although personal computer versions of IMSL and SAS are available, the authors recommend mainframe computers rather than the present generation of personal computers for speed and reliability. The currently emerging generation of personal computers probably will be able to handle the estimation with ease.

An Application of the GB2 Family

Introduction. Cummins and Freifelder [1978] studied a set of fire claims from a major university. They found that the lognormal and gamma distributions did not have sufficiently heavy tails to describe the data accurately. The log-t distribution was found to provide a better fit to the tail data. Since the objective of their article was to estimate percentile points in the tail of the total claims distribution, the t-distribution was selected since it provided the best fit to the tail of the severity distribution among the distributions tested. The total claims distribution was then estimated by simulation and using well-known approximation methods such as the normal power (NP) method.

Although the log t-distribution provided an adequate fit to the data studied by Cummins and Freifelder, it is not the optimal model for their data set. One problem is that the logs of their fire claim data have non-zero skewness (skewness = $\alpha_3^* = 1.2$), whereas the log-t is a symmetrical distribution. The purpose of this section is to illustrate the use of the GB2 family by testing other members of the family utlizing the Cummins-Freifelder fire claims data.

Data and Hypotheses. The Cummins-Freifelder data set consists of 80 fire claims. The data cover several years and have been adjusted to a common time point using a claims cost index maintained by the university from which the claims were obtained. The data are described in more detail in Cummins and Freifelder [1978].

The first step in the analysis was to plot as a histogram and to compute the

empirical *mean residual life* (MRL) (see Hall and Willner [1981] and Hogg and Klugman [1984]). The MRL is defined as follows:

$$(4) \qquad e(M) = \frac{\int_{M}^{\infty} [1 - F(x)]\, dx}{[1 - F(M)]}$$

The procedure is to compare the empirical mean residual life with the theoretical MRL of the probability distributions under consideration. Because the data are characterized by a heavy tail, natural logarithms were used. The shape of the empirical MRL function for the sample suggests a log-gamma or log-Weibull distribution (Hogg and Klugman [1984, p. 109]). Since the Weibull is a special case of the Burr, the log-Burr distribution was chosen for analysis. Thus, if the random variable x represents claim severity, the hypothesized distributions for $y = \ln x$ are as follows:[3]

$$(5) \qquad g(y) = \frac{y^{p-1} e^{-y/\beta}}{\beta^{p} \Gamma(p)}, \quad y > 0 \qquad \text{Gamma}$$

$$(6) \qquad h(y) = \frac{aqy^{a-1}}{b^{a}} [1+(y/b)^{a}]^{-(q+1)}, \quad y > 0 \qquad \text{Burr}$$

The log-t distribution also was used for purposes of comparison (see Cummins and Freifelder [1978, p. 47]).

Actually, there are two forms of the Burr -- equation (5) and an "inverse" distribution obtained by setting $z = \beta/y$. The inverse distribution extends the region of coverage of the distribution toward low kurtosis values for given skewness. "This extension of the region covers virtually all of the bell-shaped and J-shaped regions of the beta (Pearson Type I) region." (Burr[1973, p. 2].) Thus, to find the best-fitting Burr distribution, it is necessary to consider both the Burr and inverse-Burr. Both the Burr and the inverse-Burr have closed-form distribution functions.

Neither the gamma nor the Burr has a closed form MRL. However, the gamma MRL is easily obtained by evaluating incomplete gamma functions. The MRL of the Burr is given below:

$$(7) \qquad e_B(M) = \frac{\int_{M/b}^{\infty} (1+x^a)^{-q}dx}{[1+(M/b)^a]^{-q}} = \frac{\int_{0}^{[1+(M/b)^a]^{-1}} t^{q-(1/a)-1}(1-t)^{(1/a)-1}dt}{[1+(M/b)^a]^{-q}}$$

The change of variable $x = (y/b)$ has been employed in the left hand fraction. The expression in the numerator in the right-hand fraction is an incomplete Beta function, $B_z(q-1/a,1/a)$. The transformation that yields the Beta function is $s = (1+y^a)^{-1}$. The Beta function form of $e_B(M)$ is important because some computer programs include this function as a library routine.

Results. The results of the parameter estimation are presented in Table 1. The parameters of the log-gamma and log-t distributions were estimated using the method of maximum likelihood. The parameters of the Burr were estimated using a moments-based method proposed by Burr [1973] and discussed by Tadikamalla and Ramberg [1975]. The method is designed to match the first four moments of the Burr with the first four moments of the data.

The steps in the estimation of the Burr are as follows: (1) Compute the mean (μ^*), standard deviation (σ^*), skewness (α_3^*), and kurtosis (α_4^*) of the data. (2) Solve the simultaneous nonlinear equations $\alpha_3(a,q) = \alpha_3^*$ and $\alpha_4(a,q) = \alpha_4^*$ for a and q, where $\alpha_3(a,q)$ and $\alpha_4(a,q)$ represent the theoretical skewness and kurtosis of the distribution. (3) Compute the mean (μ) and standard deviation (σ) of the Burr implied by the estimates of a and q. Percentile points of the Burr can then be obtained by using the following formula $H(y^*) = 1 - [1+(z\sigma+\mu)^a]^{-q}$, where $z = (y^*-\mu^*)/\sigma^*$ and y^* = the log of observed severity. Thus, the mean, variance, skewness, and kurtosis of $h(y^*)$ will match the corresponding moments of the data.

The only part of the estimation procedure that presents any difficulties is step (2), which involves solving nonlinear simultaneous equations. The authors obtained these solutions using IMSL. However, for many cases, sufficiently accurate estimates can be obtained by using the tables presented in Burr [1973].

Since the estimation technique for the Burr is based on a standardization

procedure, some values of the observed sample may not be within the feasible range of the estimated distribution. The requirement is that $z\sigma+\mu \geq 0$ or that $y^* \geq \mu^*-\mu\sigma^*/\sigma$. There is no guarantee that all of the sample values will satisfy this requirement. If the number of observations outside of the feasible range is substantial, it is clear that the Burr does not model the data accurately. However, if only a small proportion of the observations falls into this category and if the distribution provides an accurate fit in the area of the curve of most interest in the analysis (e.g., the right hand tail), then it may be reasonable to use the distribution.

The empirical skewness (α_3^*) and kurtosis (α_4^*) indicate that the data fall very close to the boundary line between the Burr and the inverse-Burr. Both distributions were tested, and the Burr proved to give the better fit. Only the smallest observation falls outside of the feasible range of the estimated distribution.

The Chi-square statistics shown in Table 1 indicate that the log-Burr and log-gamma both provide a better fit to the data than the log-t. However, the chi-square results for the log-Burr and log-gamma are quite comparable and do not support the choice of one distribution over the other. This is especially true in view of the fact that the chi-square is not a very powerful test for continuous distributions.

To provide more information on the fit of the candidate distributions, the distribution functions and their tails are plotted in Figures 2 and 3, respectively. The results indicate that the log-Burr provides a slightly better fit over the entire range of observed data, even though the first observation is below the minimum permissable value for the fitted distribution. The log-Burr also provides a better fit in the tail (see Figure 3), which is represented here by the last fifteen observations. Similar conclusions are supported by the plot of fitted and empirical MRLs presented in Figure 4.

Implications. The consequences of using the lighter-tailed log-gamma distribution rather than the log-Burr are potentially quite serious. To provide an

example, we calculated the expected value of the reinsurer's severity distribution under an excess of loss contract. The reinsurer's obligation is given below:

$$(8) \qquad X_{RE} = \begin{cases} 0 & \text{for } X \leq M \\ X - M & \text{for } M < X < U \\ U & \text{for } X \geq U \end{cases}$$

where X_{RE} = the reinsurer's obligation,

X = loss amount,

M, U = lower and upper boundaries of the reinsurance layer.

Since logarithmic distributions are used, the reinsurance severity expectations are given by the following formula:

$$(9) \qquad E(X_{RE}) = \int_{M}^{U} e^{X} \, dF(X) + e^{U} [1-F(U)] - e^{M} [1-F(M)]$$

To obtain the reinsurance pure premium, $E(X_{RE})$ would be multiplied by expected claim frequency.

For illustrative purposes, the upper boundary was set at ln U = 14 and the lower boundary allowed to vary between ln M = 7 and ln M = 14. The upper boundary point was chosen to be slightly higher than the largest observed claim ($\ln X_{MAX} = 13.35$). The results are graphed in Figure 5. The reinsurance premiums implied by the log-Burr are clearly substantially higher than those implied by the log-gamma, and the discrepancy would be greater if the upper boundary had been set higher.

The choice of severity distribution also has a substantial effect on the total claims distribution. This distribution was estimated by simulation. The experiment involved selecting 10,000 draws from a negative binomial frequency distribution with failure parameter 0.03714 and shape parameter 95.1968, followed by the appropriate number of random draws from the fitted severity distributions. The frequency distribution is the one utilized by Cummins and Freifelder [1978] in their analysis of the fire claim sample. The 99th percentiles of the simulated total claims distributions are 265,817 for log-gamma severity and 792,466 for log-Burr severity. Although this can partly be explained by the fact that the fitted log-gamma has a

lower minimum value than the log-Burr, most of the discrepancy is attributable to the heavier tail of the latter distribution.[4]

Conclusions

This article proposes the GB2 family of probability distributions for use in modeling loss severity in insurance. Because of the heavy tails characterizing many insurance claim processes, modelling of the log-transform of severity will often be appropriate. The advantage of the GB2 family is that it has a maximum of four parameters and thus can be fitted to a wide range of processes observed in practice. Most of the distributions traditionally applied in the risk theory literature are members of the family. These include the lognormal, the Weibull, the gamma, and the generalized gamma. Estimation of the more complicated members of the GB2 family requires the solution of nonlinear simultaneous equations. However, because of advances in computing, this is not a serious impediment.

Two members of the GB2 distribution were fitted to a set of eighty fire claims. The results indicate that the log-Burr distribution provided an excellent fit to the data. The log-gamma distribution also did well on an overall basis. However, the tail of the latter distribution proved to be lighter than the tail of the observed data. Utilizing a lighter tailed severity distribution can have a magnified effect on reinsurance premiums and the estimated distribution of total claims, especially when logarithmic probability distributions are used. Thus, readers are advised to supplement the usual goodness of fit tests with a careful examination of the tail behavior of the observed and fitted distributions. The mean residual life methodology is an excellent approach to gauging the fit in the tail.

Table 1

Moments and Parameter Estimates for Fire Loss Data

Distribution	Estimation Method	Parameters		X^2-Statistic
Log-Student's t	Maximum Likelihood	$m = 8.0159$	$w = 0.9455$	15.50
		$\gamma = 4.4832$		
Log-gamma	Maximum Likelihood	$p = 40.5717$	$b = 0.20248$	6.50
Log-Burr	Moments	$a = 24.6293$	$q = 1.5045$	5.75
		$\mu = 0.1097$	$\sigma = 0.0765$	
Observed Data (Logs):		Mean = $\mu^* = 8.2151$	Standard Deviation = $\sigma^* = 1.3490$	
		Skewness = $\alpha_3^* = 1.2067$	Kurtosis = $\alpha_4^* = 4.8753$	

Note: The three-parameter Student's t distribution is given in Cummins and Freifelder [1978]. For the Chi-square test, the method of intervals with equal probability was used (Kendall and Stuart [1977]). The hypothesis that the data came from a log-gamma or a log-Burr can be accepted with $\alpha = .5$.

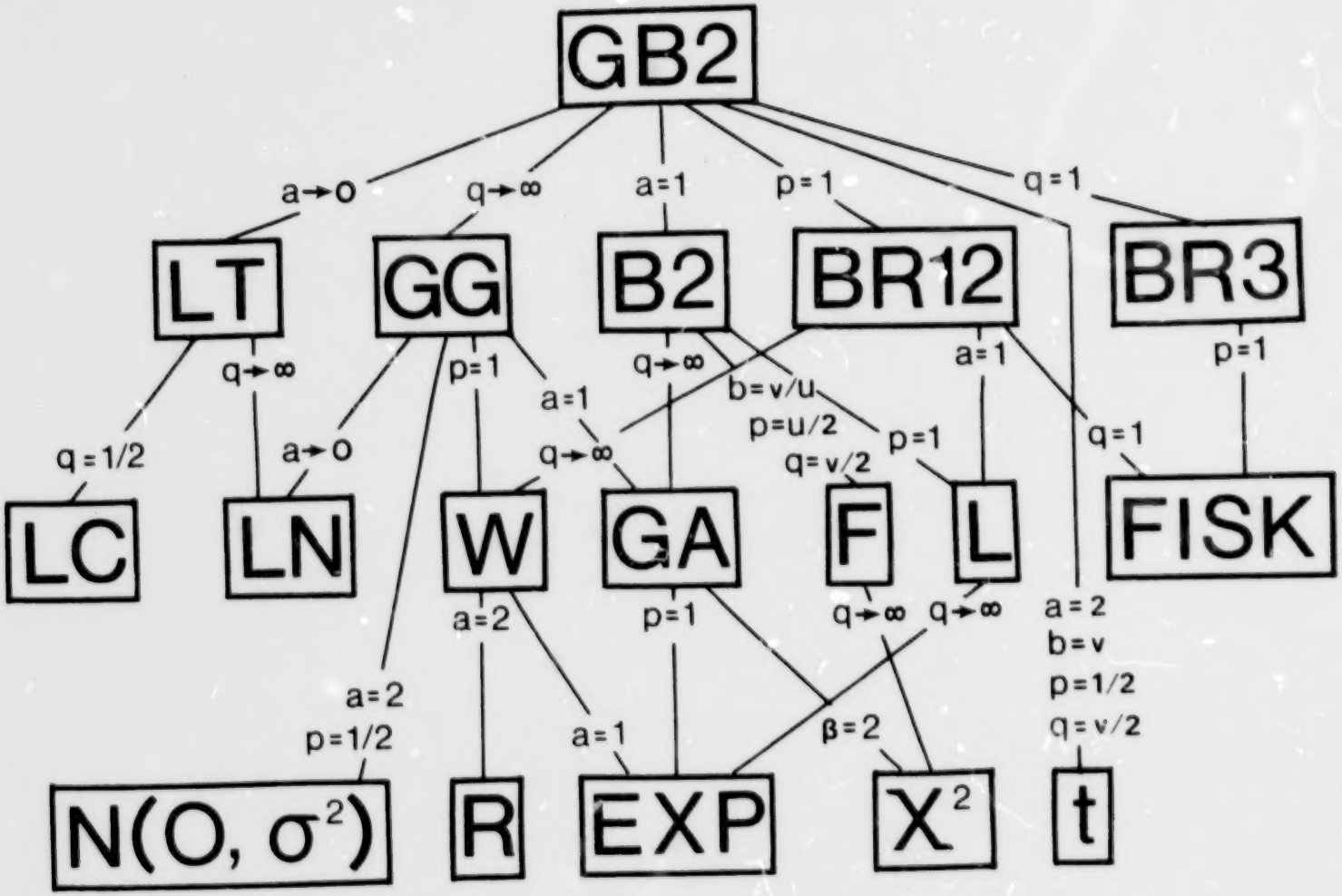


Figure 1
Generalized Beta: Type 2
Distribution Family Tree

Source: McDonald & Butler (1985).

FIG 2: DATA V. FITTED CDFs

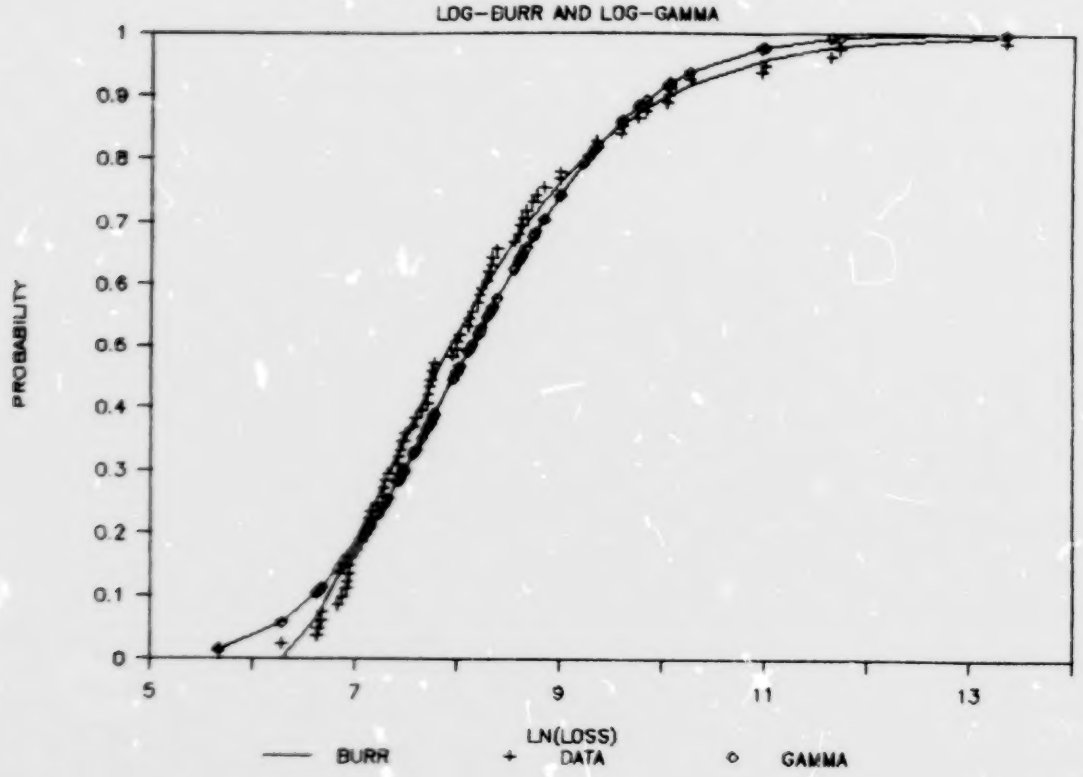


FIG 3: DATA V. FITTED TAILS

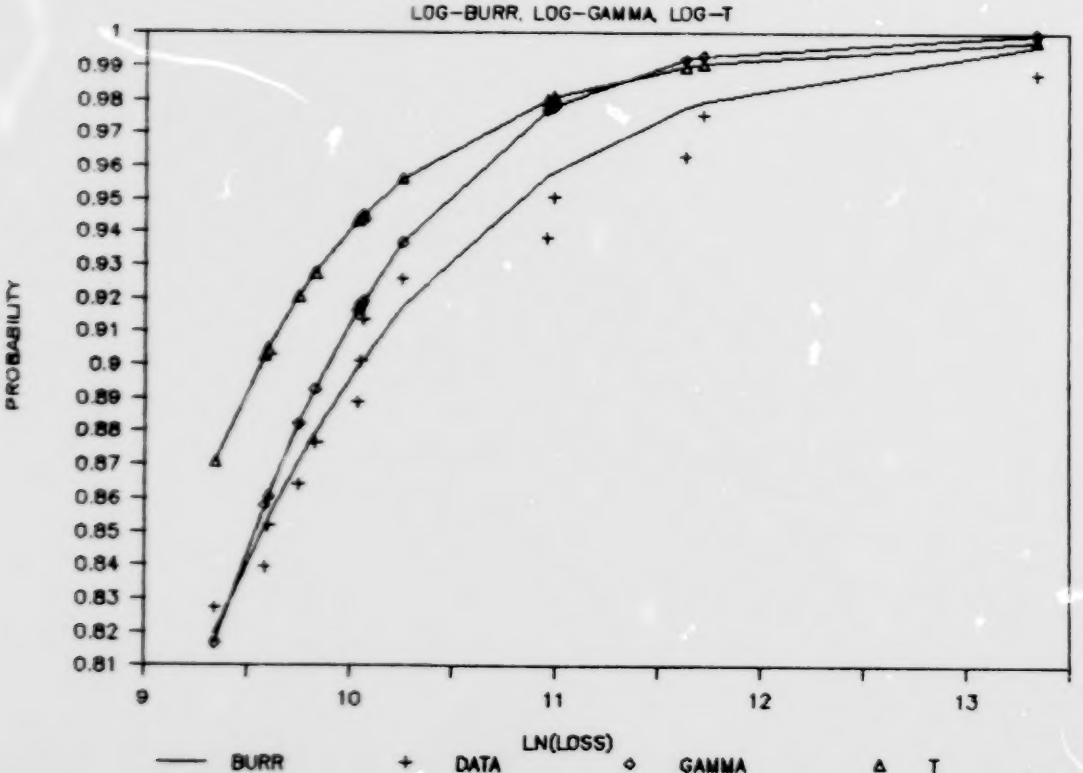


FIG 4: MEAN RESIDUAL LIFE

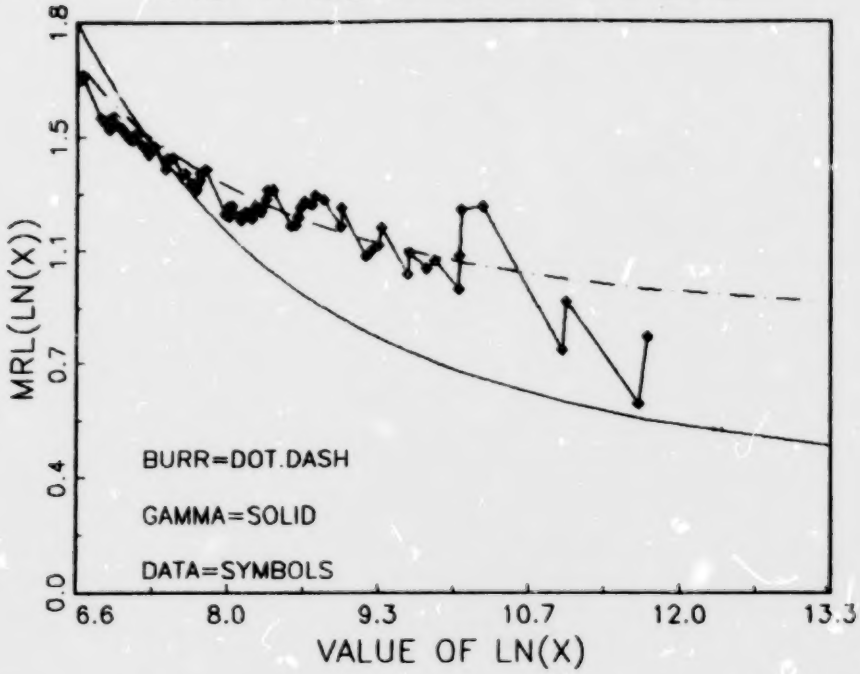
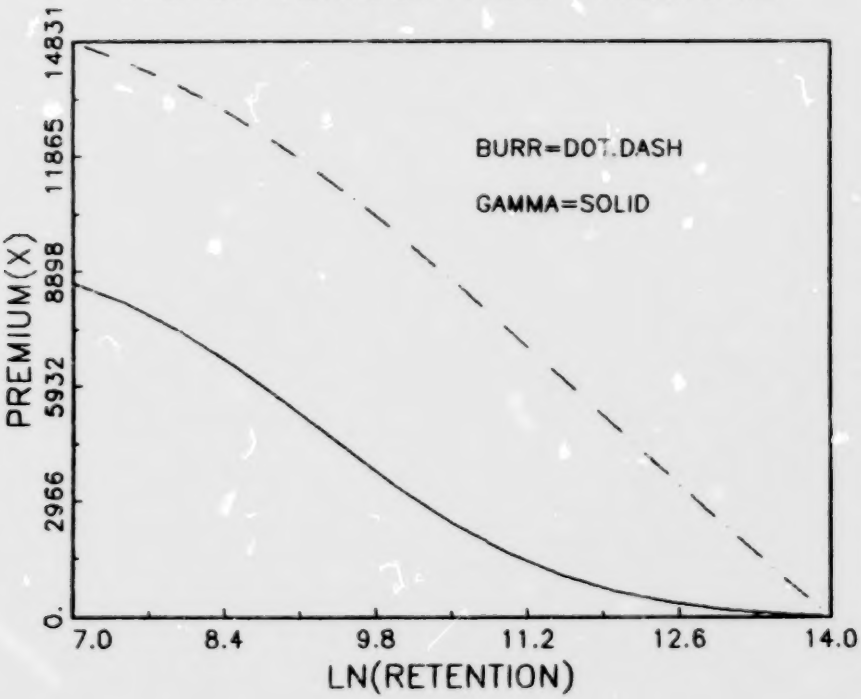


FIG 5: REINSURANCE PREMIUMS



FOOTNOTES

[1]The GB1 distribution can be obtained from the GB2 by making the transformation $w = x/[1+(x/b)^a]^{1/a}$. If x is GB2, w will be GB1.

[2]The following abbreviations are used in Figure 2: GB2 = generalized beta of type II, LT = log-t, LC = log-Cauchy, GG = generalized gamma, LN = lognormal, N = normal, W = Weibull, R = Rayleigh, B2 = beta of the second kind, GA = gamma, EXP = exponential, BR12 = Burr type 12 (given in equation (6) below), F = F-distribution, t = student's t, χ^2 = chi-square, and BR3 = Burr type 3.

[3]The parameter β in the gamma distribution is equal to related to the GB2 parameters as follows: $b^a = q\beta^a$.

[4]Tests were also conducted using the three-parameter gamma distribution, where $X \geq \delta$, $\delta > 0$ (see Johnson and Kotz [1970, p. 166]). However, the two parameter gamma distribution provided a better overall fit to the fire claim sample.

REFERENCES

Abramowitz, Milton and Irene A. Stegun [1965]. *Handbook of Mathematical Functions.* New York: Dover Publications.

Aiuppa, Thomas A. [1986]. "Maximum Probable Yearly Aggregate Loss: A Pearson Approximation." Forthcoming in *Journal of Risk and Insurance.*

Beard, R. E., T. Pentikainen, and E. Pesonen [1984]. *Risk Theory.* 3d ed. New York: Chapman and Hall.

Buhlmann, Hans [1970]. *Mathematical Methods In Risk Theory.* New York: Springer- Verlag.

Burr, Irving W. [1973]. "Parameters for a General System of Distributions To Match a Grid of α_3 and α_4." *Communications In Statistics* 2: 1-21.

_____ and Peter J. Cislak [1968]. "On General Systems of Distributions." *Journal of the American Statistical Association* 63: 627-635.

Cummins, J. David and Leonard R. Freifelder [1978]. "A Comparative Analysis of Alternative Maximum Probable Yearly Aggregate Loss Estimators." *Journal of Risk and Insurance* 45 (March): 27-52.

Edlefson, Lee E. and Samuel D. Jones [1986]. *Gauss.* Kent, WA: Applied Technical Systems.

Elderton, W.P. and N. L. Johnson [1969]. *Systems of Frequency Curves.* Cambridge: The University Press.

Hall, W. J. and J. A. Willner [1981]. "Mean Residual Life." In M. Csorgo, et al., eds., *Statistics and Related Topics.* Amsterdam: North-Holland Publishing Co.

Hogg, Robert V. and Stewart A. Klugman [1984]. *Loss Distributions.* New York: John Wiley & Sons.

IMSL [1985]. *International Mathematics and Statistics Library.* Houston, TX.

Johnson, Norman L. and Samuel Kotz [1970]. *Continuous Univariate Distributions -- 1.* New York: John Wiley & Sons.

Kendall, Maurice C. and Alan Stuart [1977]. *The Advanced Theory of Statistics.* 4th ed. New York: Macmillan Publishing Co.

McDonald, James B. [1984]. "Some Generalized Functions for the Size Distribution of Income." *Econometrica* 52 (May): 647-663.

_____ and Richard J. Butler [1985]. "Some Generalized Mixture Distributions." Working Paper. Brigham Young University, Provo, Utah.

Paulson, A.S. and N.J. Faris [1985]. "A Practical Approach to Measuring the Distribution of Total Annual Claims." In J. D. Cummins, ed. *Strategic Planning and Modeling In Property-Liability Insurance.* Hingham, MA: Kluwer Academic Publishers.

SAS Institute [1985]. *SAS User's Guide: Statistics.* Version 5 ed. Cary, NC.

Tadikamalla, Pandu R. and John S. Ramberg [1975]. "An Approximate Method for Generating Gamma and Other Variates." *Journal of Statistical Computation and Simulation* 3: 275-282.